DOCUMENT RESUME

ED 246 071 TM 840 342

AUTHOR O'Brien, Michael L.; Tohn, Diane
TITLE Applying and Evaluating Rasch Vertical Equating Procedures for Out-of-Level Testing.
PUB DATE 10 Feb 84
NOTE 25p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (West Palm Beach, FL, February 10, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Elementary Education; *Equated Scores; *Latent Trait Theory; Reading Comprehension; Statistical Analysis
IDENTIFIERS *Equipercentile Equating; *Out of Level Testing; Rasch Model; Vertical Equating

ABSTRACT

This study investigated the development and application of two widely used methods of vertical equating: Rasch and Equipercentile. The assumptions underlying both techniques are discussed. Procedures for utilization of the Rasch method are illustrated. The study was conducted to determine whether, based upon Rasch vertical equating, a local school district should administer out-of-level tests (tests not on actual grade level) to exceptionally able students. A comparison was made between the school district's equating results and that of the test publisher's vertically scaled scores based on Equipercentile equating. Results indicated the publisher's vertical scale was comparable to the scale estimated from the local school district through use of Rasch equating. Limitations of both methods are discussed. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED246071

APPLYING AND EVALUATING RASCH VERTICAL EQUATING PROCEDURES FOR OUT-OF-LEVEL TESTING

Michael L. O'Brien and Diane Tohn

Teachers College, Columbia University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. L. O'Brien

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the Eastern Educational Research Association, West Palm Beach, Florida, February 10, 1984.

TM 810 342

ABSTRACT

This study focused on the development and application of two widely used methods of vertical equating-- Rasch and Equipercentile. The assumptions underlying these techniques are discussed and the procedures for using the Rasch method are illustrated. The purpose of the study was to determine whether, based upon Rasch vertical equating, a local school district should administer out-of-level tests to exceptionally able students. The results showed that the test publisher's vertical scale, developed by the Equipercentile method, was comparable to the scale estimated from the local school district via Rasch equating. Limitations of the methods are also discussed.

While it is of great interest to scale measures vertically for across grade (eg. growth) information, it is particularly important to have vertical scales for out-of-level testing problems. Theoretically, if a student is given a test of "appropriate difficulty," the standard error of measurement will be minimal. Out-of-level testing may be used to facilitate an optimal, "higher-level" measure of a student's achievement in place of the "on-level" test that would be given to that student according to his or her actual grade level. Thus, for an unusually high achieving student, administering a higher level of a test should provide a more accurate estimate of his or her educational performance. However, the realization of a set of vertically equated tests relies upon careful planning and confirmation. In order for tests to be vertically equated, item content and difficulty must be closely examined so that the vertical scale presents a true picture of the progress of a student in relation to a wide range of educational achievement.

THEORETICAL FRAMEWORK

Although the concept of out-of-level testing is highly regarded by practitioners, the validity of its use depends upon the "success" of the method used to vertically equate the levels of an achievement measure. In order to evaluate such quality, one must determine the extent to which the assumptions of a given method are met.

In order to equate measures vertically using the Equipercentile method, adjacent test levels must be administered to the same sample of students. Theoretically, the students would retain the same relative ability level on each test; however, each person would be likely to answer fewer items correctly on the higher (ie. harder) level test. Thus, since each person takes both test levels, there are only three basic assumptions underlying Equipercentile

equating. First, the test levels must measure the same underlying trait. Second, careful random sampling of examinees must be made since the equating is population specific. Third, each test level must be reliable so that the standard error of measurement is minimal for the "target" ability of each level. While these assumptions can be easily met in most situations, the main disadvantage of Equipercentile equating is inefficiency. This method is expensive to use and taxing on the students who must take two tests.

A more practical procedure for vertical equating is the latent trait Rasch method. Compared to the traditional (eg. Equipercentile) methods, latent trait equating procedures are more efficient and less dependent upon sampling characteristics. While there are three possible latent trait models useful for vertical equating, the Rasch model is the easiest to apply and has been shown to be the most consistent (Sontag, 1984). Further, only the Rasch model may be applied to small sample sizes. In order to use the Rasch method of vertical equating, a set of common, "anchor" items are administered to <u>different</u> samples of persons. Thus, vertical equating can be performed based upon each person taking <u>only</u> the test level that is appropriate. Test levels are then equated from the average, relative change in the difficulties of the anchor items. The basic, underlying assumption of the Rasch method is that there is a <u>single constant</u>, G_{AB}, necessary to translate trait estimates along two levels of a vertical scale, which is invariant with respect to the scale values of the anchor item <u>within</u> either vertical form.

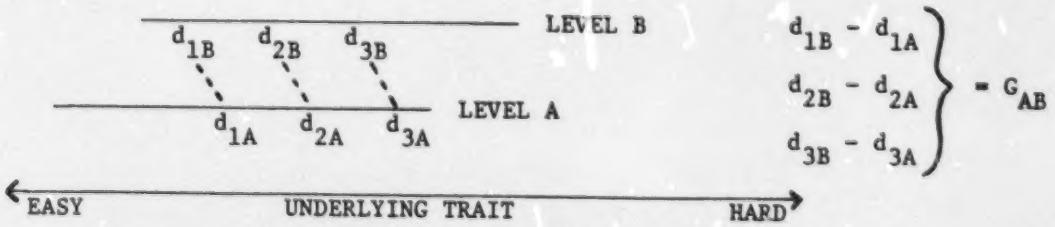

Fig. 1.--Schematic of the Rasch method of vertical equating.

From Figure 1, it can be seen that, given three anchor items, d_1, d_2, and d_3, the relative "shift" in their respective difficulties between Levels A and B is assumed to be the translation constant G_{AB}. That is, anchor items are assumed to retain the same relative difficulty when administered to samples of persons who vary in ability. The specific assumptions of the Rasch method require careful test development and analysis procedures. First, the anchor items must have been experienced by all persons. While anchor items would obviously tap the same trait, the examinees at both levels must be familiar with the content of these items so that chance level does not become a factor. This first assumption may be tested using item fit statistics as discussed by Wright and Stone (1979). Second, since the equating is based only upon the set of anchor items, or "anchor test," it is necessary that the content of the anchor test be comparable to the total test at any given level. Similarly, a representative range of item difficulty should be built into the anchor test to properly equate across ability levels. This assumption may be met in the construction of vertical scales. Third, the tests should be administered to a representative sample of persons. While Rasch procedures are, theoretically, "sample independent," experience shows that representative sampling never hurts! It is particularly important that each test level be given to an appropriate ability group. Thus, in terms of Rasch procedures, optimal results are attained when the mean ability, $\bar{b}$, of a sample is close to zero.

Whereas previous research focused on comparing vertical equating methods against a criterion (Holland & Rubin, 1982; Cook & Douglass, 1982; Slinde & Linn, 1979), their inconsistent findings point to a need to focus on the evaluation of assumptions of the individual methods. Studies comparing Equipercentile v. Rasch methods of vertical equating found that the two methods

yielded similar results (Rentz & Lenke, 1982; Guskey, 1979). These results would suggest that the more practical Rasch procedure would be preferable if the assumptions of the method could be met.

PROBLEM

Out-of-level testing is used to facilitate a more reliable estimate of a person's ability by providing a test of appropriate difficulty to measure that person. However, the measure will be accurate on the vertical scale only to the extent that the vertical equation meets the assumptions of the equating methods. Moreover, when a test publisher supplies vertically equivalent scores based upon the Equipercentile equating method, it is of interest to a local school district to verify the applicability of such measures to the resident students. By applying and verifying the Rasch method of vertical equating to the local test data, it is possible to determine how closely the vertically equated scores based on the local students agree with the test publisher's table of equivalent scores. That is, does the benefit of increased measurement efficiency of the out-of-level test outweigh any possible bias in the student's vertically scaled score?

METHODS

Sample and Instrumentation

The problem stemmed from a sample of 459 sixth graders enrolled in a suburban New York school district during the 1982-1983 school year. Since approximately one-third of these students' reading comprehension ranked at or above the 97th percentile nationally, administering an out-of-level test was clearly advisable.

Levels H and I of the 50-item Sequential Tests of Educational Progress (STEP) were used to measure the students' reading comprehension. These tests were published by Educational Testing Service (ETS) and were vertically equated

by the publisher through the use of the Equipercentile method. Both levels evidenced sufficient validity and reliability and were widely used for measuring reading comprehension. The target grade range of Level H was 6.5 - 7.5 (ie. "on-level" for sixth graders); for Level I, 7.5 - 10.5 (ie. "out-of-level" for sixth graders).

Procedures

The students took either Level H (N=121) or Level I (N=238) according to the test deemed more appropriate for each student by his or her teacher. Since each student took only one level of the STEP test, only the Rasch method could be used to vertically equate Levels H and I for these students. A random sample of 100 students from each test was drawn using a table of random numbers. According to Wright and Stone (1979), the Rasch vertical equating procedure may be usefully applied to a sample of 100 persons taking as few as 10 anchor items. The student responses were then analyzed using the BICAL computer program (Wright & Mead, 1977). The 15 anchor items linking Levels H and I were subsequently investigated by the following methods as given by Wright and Stone.

Unlike the Equipercentile method, with the Rasch model it is possible to evaluate the quality of the vertical equation by inspecting anchor items collectively and individually. The stability of the assumed translation constant, G_{HI}, necessary to equate Levels H and I can be tested <u>across</u> the set of anchor items by a chi-square test (1) with χ^2_k :

$$\sum_{i}^{k} (d_{i_H} - d_{i_I} - G_{HI})^2 \quad (N/12)\,[k/(k-1)] \sim \chi^2_k \qquad (1)$$

wherein: $\sum_{i}^{k}$ is the summation across item i to k items;

d_{i_H} is the Rasch difficulty of anchor item i in Level H;

d_{i_I} is the Rasch difficulty of anchor item i in Level I;

G_{HI} is the mean difference between Rasch difficulties of the k pairs of anchor items;

N is the number of persons taking each form; and

k is the number of anchor item pairs.

The stability of individual anchor items can be evaluated by (2) with χ^2_1:

$$(d_{i_H} - d_{i_I} - G_{HI})^2 \quad (N/12) \quad [k/(k-1)] \sim \chi^2_1 \qquad (2).$$

If the χ^2_k were statistically significant beyond α=.05, any individual pair of anchor items with a χ^2_1 beyond α=.05 would be deleted. The set of k "good" anchor items would determine the G_{HI} necessary to equate Rasch ability estimates for any given Total Raw Score on Level I to its equivalent score on Level H. It would now be possible to compare the observed vertically equated Total Percent Correct Score estimated from the Rasch method to the expected scores supplied by the test publisher (via the Equipercentile method) through the use of a chi-square test of association.

RESULTS AND DISCUSSION

Applying and Evaluating Rasch Procedures

The original 15-item anchor test. The results from the analysis of the original, 15-item anchor test are summarized in Table 1. The sequence number of each anchor item along with its content category can be seen in the column labelled "Item Name H/I, Content $_{HI}$." The Rasch item difficulties and traditional item P-values for the anchor items are reported in the columns labelled "d_H (P_H)" and "d_I (P_I)," respectively. Low Rasch item difficulties indicate relatively easy items; high d_i's correspond to harder items. Item discrimination estimates are shown in the "$disc_H$" and "$disc_I$" columns; Rasch item fit

TABLE 1.-- Summary data from the Rasch analysis of the original, 15-item anchor test

ITEM NAME H/I CONTENT$_{HI}$	d_H	(P_H)	$disc_H$	TOTAL FIT-T_H	d_I	(P_I)	$disc_I$	TOTAL FIT-T_I	χ^2_1	
37/2 LC[1]	0.06	(77)	1.18	-0.90	0.70	(71)	-0.16	2.57	13.56	(.01)
36/1 VC[2]	0.13	(76)	0.52	1.35	-2.02	(97)	0.55	0.10	22.46	(.01)
49/14 IC[3]	0.13	(76)	1.34	-0.93	-0.07	(83)	1.48	-0.83	1.32	(NS)
39/4 IC	0.44	(71)	1.53	-1.70	-2.43	(98)	0.93	0.12	47.36	(.01)
42/7 LC	0.77	(65)	0.90	0.36	0.22	(79)	1.05	0.02	0.01	(NS)
47/12 LC	2.00	(40)	0.68	1.50	1.89	(47)	0.65	1.11	2.00	(NS)
46/11 VC	1.47	(51)	0.37	1.97	0.47	(75)	0.64	0.66	1.61	(NS)
38/3 LC	0.13	(76)	0.98	-0.40	-2.02	(97)	1.69	-0.06	22.46	(.01)
40/5 IC	-0.40	(83)	1.39	-0.62	0.15	(80)	0.95	0.27	11.63	(.01)
41/6 LC	-0.01	(78)	1.21	-0.70	-0.15	(84)	0.71	0.47	1.76	(NS)
43/8 IC	0.44	(71)	1.82	-2.86	-0.15	(84)	0.76	0.09	0.00	(NS)
44/9 IC	0.93	(62)	1.17	-0.73	1.79	(49)	0.39	2.15	18.89	(.01)
50/15 IC	1.76	(45)	1.02	-0.54	1.70	(51)	1.04	-0.03	2.46	(NS)
45/10 IC	1.28	(55)	0.99	0.15	3.28	(22)	0.71	0.51	60.64	(.01)
48/13 IC	2.00	(40)	0.93	0.50	-0.97	(92)	1.05	-0.22	52.04	(.01)
	$\bar{d}_H$= 0.74		$\overline{disc}_H$=1.07	$\overline{Fit\text{-}T}_H$=-0.24	$\bar{d}_I$ =0.16		$\overline{disc}_I$=0.83	$\overline{Fit\text{-}T}_I$=0.46		
	SD_H=0.89		SD_H=0.61	SD_H= 1.12	SD_I=1.27		SD_I=0.66	SD_I=0.94		

G_{HI} = 0.58 (overall χ^2 = 258.20 ($p < .01$)

$r_{d_{HI}}$ = 0.46 ($p < .10$); $r_{disc_{HI}}$ = 0.11 (NS); $r_{Fit\text{-}T_{HI}}$ = 0.04 (NS); $\bar{b}_H$ = 1.46; $\bar{b}_I$ = 1.74; $KR20_H$=0.91; $KR20_I$=0.82

1 "LC" indicates Literal Comprehension; 2 "VC" indicates Vocabulary; 3 "IC" indicates Inferential Comprehension

statistics appear in the "Total Fit-T_H" and "Total Fit-T_I" columns. Finally, the χ^2_1 statistic evaluating the fit of each $d_H - d_I$ against G_{HI} is given in the "χ^2_1" column, along with its corresponding significance level ($|\chi^2_{1 \cdot 95}|=3.8$; $|\chi^2_{1 \cdot 99}|=6.6$)

From Table 1 it can be seen that the overall χ^2_{15} was statistically significant beyond $\alpha=.01$ ($\chi^2_{15}=258.20$), indicating that the G_{HI} = 0.58 logits was inconsistent across the anchor items. Thus, the original anchor test did not meet the main assumption of the Rasch vertical equating method. This was not surprising as one scans the item difficulty estimates for several items that were actually harder at the higher Level I (Items 45/10, 37/2, 40/5, and 44/9)! Two of these items concurrently misfit the Rasch model (T>2.0), indicating that lack of unidimensionality can explain problems of faulty anchor items. Item 10 of Level I appeared to have been miskeyed. Items 36/1, 39/4, 38/3, and 48/13 were relatively too easy at Level I. Overall, the correlation between the anchor item difficulties was 0.46 ($p < .10$), indicating the disproportionate shift of the relative difficulties between Levels H and I. Similarly, the correlations between item discrimination estimates and Rasch item fit statistics, respectively, were of zero order ($r_{disc_{HI}}=0.11$; $r_{Fit\text{-}T_{HI}}=0.04$). This result indicated that fitting additional item parameters would be unlikely to improve the overall poor quality of the anchor test.

The revised seven-item anchor test. The results from the original anchor test indicated the need for the evaluation of assumptions before proceeding with the actual equating. Despite the many problems seen with the original anchor items, seven items appeared to form a good, revised anchor test. The summary data for this revised set of items are shown in Table 2. From Table 2 it can be seen that the overall χ^2_7 was not statistically significant ($\chi^2_7=7.04$), indicating that the G_{HI} = 0.38 logits was a stable translation

TABLE 2.-- Summary data from the Rasch analysis of the revised, seven-item anchor test

ITEM NAME H/I CONTENT$_{HI}$	d_H	(P_H)	$disc_H$	TOTAL FIT-T_H	d_I	(P_I)	$disc_I$	TOTAL FIT-T_I	x_1^2	
37/2 LC[1]	----	----	----	----	----	----	----	----	----	
36/1 VC[2]	----	----	----	----	----	----	----	----	----	
49/14 IC[3]	0.13	(76)	1.34	-0.93	-0.07	(83)	1.48	-0.83	0.32	(NS)
39/4 IC	----	----	----	----	----	----	----	----	----	
42/7 LC	0.77	(65)	0.90	0.36	0.22	(79)	1.05	0.02	0.29	(NS)
47/12 LC	2.00	(40)	0.68	1.50	1.89	(47)	0.65	1.11	0.72	(NS)
46/11 VC	1.47	(51)	0.37	1.97	0.47	(75)	0.64	0.66	3.77	(NS)
38/3 LC	----	----	----	----	----	----	----	----	----	
40/5 IC	----	----	----	----	----	----	----	----	----	
41/6 LC	-0.01	(78)	1.21	-0.70	-0.15	(84)	0.71	0.47	0.57	(NS)
43/8 IC	0.44	(71)	1.82	-2.86	-0.15	(84)	0.76	0.09	0.44	(NS)
44/9 IC	----	----	----	----	----	----	----	----	----	
50/15 IC	1.76	(45)	1.02	-0.54	1.70	(51)	1.02	1.04	1.02	(NS)
45/10 IC	----	----	----	----	----	----	----	----	----	
48/13 IC	----	----	----	----	----	----	----	----	----	
	$\overline{d}_H$= 0.94		$\overline{disc}_H$=1.05	$\overline{Fit\text{-}T}_H$=-0.17	$\overline{d}_I$ =0.56		$\overline{disc}_I$=0.90	$\overline{Fit\text{-}T}_I$=0.21		
	SD_H=0.90		SD_H=0.69	SD_H= 1.28	SD_I=0.94		SD_I=0.55	SD_I=0.78		

G_{HI} = 0.38 (overall x^2 = 7.04, NS)

$r_{d_{HI}}$ = 0.92 ($p < .01$); $r_{disc_{HI}}$ = 0.57 (NS); $r_{Fit\text{-}T_{HI}}$ = 0.32 (NS); $\overline{b}_H$ = 1.46; $\overline{b}_I$ = 1.74; $KR20_H$=0.91; $KR20_I$=0.82

1 "LC" indicates Literal Comprehension; 2 "VC" indicates Vocabulary; 3 "IC" indicates Inferential Comprehension

constant across the set of revised anchor items. Thus, the revised anchor test met the main assumption of the Rasch vertical equating method. The individual pairs of anchor items each fit the G_{HI}, as can be seen from the χ^2_1 column. The positive effect of the editing procedure from the original test was best seen from the statistically significant (α=.05) rise in the correlation between the 7-item revised anchor test compared to that of the 15-item original test (r_{HI_7}=0.92, $p < .01$; $r_{HI_{15}}$=0.46, $p < .10$; one-tailed $z=1.89$, $p < .05$). The items showed acceptable Rasch model fit to their respective levels, indicating that they measured a single dimension. It was interesting to note that, as with the original anchor test, the correlations between item discrimination and Rasch Fit-T estimates, respectively, were not statistically significant.

Further assumptions using Rasch equating concern representativeness of items and persons. From Table 2, it can be seen that the range of item difficulty for each level was about two logits, which was reasonable given practical restrictions of building vertical measures. The tests were somewhat easy for the persons, which was also reasonable for an achievement test. Since a random sample from the school district population was drawn for the analysis, the persons measured were obviously representative of the group to which the analysis would be generalized. A final but important consideration was the representativeness of item content within the anchor test in relation to the total test. The comparative distribution of item content of the total test, original anchor test, and revised anchor test are presented in Table 3.

As can be seen from Table 3, the content of the revised anchor test was proportionately comparable to that of the total test. The comparability of item content between the original and revised anchor tests suggested that the problems with the poor anchor items were unrelated to item content.

TABLE 3.-- Comparative distribution of item content of total test, original anchor test, and revised anchor test

CONTENT AREA	(1) TOTAL TEST		(2) ORIGINAL ANCHOR TEST		(3) REVISED ANCHOR TEST	
	N_{ITEMS}	$\%_{TOTAL}$	N_{ITEMS}	$\%_{TOTAL}$	N_{ITEMS}	$\%_{TOTAL}$
VOCABULARY	16	32	2	13.3	1	14.29
LITERAL COMP.	16	32	5	33.3	3	42.86
INFERENTIAL COMP.	18	36	8	53.3	3	42.86
TOTAL	50	100	15	100.0	7	100.00

(1)(2) $x_2^2 = 2.90$ (NS)

(2)(3) $x_2^2 = 0.20$ (NS)

(1)(3) $x_2^2 = 1.04$ (NS)

Since an anchor test has met the assumptions of the Rasch vertical equating method, it was possible to conduct the equating and to compare the results to the publisher's Equipercentile table.

Comparison of Rasch v. Equipercentile Methods

Performing the Rasch equating. Given the Rasch ability estimates for the Total Test on Levels H and I, respectively, the Total Raw Score on Level I can be expressed in terms of an equated Level H score by adding G_{HI} (0.38 logits) to each Rasch ability estimate from Level I. One might think of this as the "handicap" for a student who took the more difficult level that measures a relatively higher level of achievement. For example, any given student whose Total Raw Score, say, of 40 on Level I would be equivalent to a Total Raw Score of about 47 on Level H. Thus, the student would have to answer seven more items correctly on the lower Level H in order to be of equivalent ability as the student taking the higher Level I.

The results of the Rasch vertical equating are presented in Table 4. The Rasch ability estimates for each Total Raw Score on Levels H and I, respectively, are shown in the columns labelled "Rasch b_H" and "Rasch b_I." The equated Rasch ability estimates putting Level I abilities in terms of Level H are reported in the "Rasch $b_{I \leftarrow H}$" column (0.38 logits were added to each Level I estimate). Finally, the equated Total Raw Score of Level I in terms of Level H are given in the "Equated I←H Rasch" column. These estimates were derived from a linear interpolation of the Total Raw Score based on each Rasch $b_{I \leftarrow H}$ v. Rasch b_H. For example, From Table 4 it can be seen that the Rasch $b_{I \leftarrow H}$ for a Total Raw Score of 41 was 2.32 logits, which corresponded to the same Rasch b_H ability estimate for a Total Raw Score of 44. That is, based upon the Rasch equating, a Total Raw Score of 41 on Level I was equivalent to a Total Raw Score of 44 on Level H.

TABLE 4.-- Results of the Rasch vertical equating

TOTAL RAW SCORE	RASCH b_H	RASCH b_I	RASCH $b_{I \leftarrow H}$	EQUATED I←H RASCH
50	---	---	---	---
49	4.28	4.56	4.94	---
48	3.56	3.81	4.19	48.88
47	3.12	3.35	3.73	48.24
46	2.80	3.01	3.39	47.61
45	2.54	2.74	3.12	47.00
44	2.32	2.50	2.88	46.25
43	2.13	2.29	2.67	45.50
42	1.95	2.11	2.49	44.77
41	1.79	1.94	2.32	44.00
40	1.65	1.78	2.16	43.16
39	1.51	1.64	2.02	42.39
38	1.38	1.50	1.88	41.56
37	1.26	1.37	1.75	40.71
36	1.14	1.24	1.62	39.79
35	1.03	1.12	1.50	38.92
34	0.92	1.00	1.38	38.00
33	0.81	0.88	1.26	37.00
32	0.71	0.77	1.15	36.08
31	0.61	0.66	1.04	35.09
30	0.51	0.55	0.93	34.09
29	0.41	0.45	0.83	33.18
28	0.31	0.34	0.72	32.10
27	0.21	0.23	0.61	31.00
26	0.11	0.13	0.51	30.00
25	0.02	0.03	0.41	29.00
24	-0.08	-0.08	0.30	27.90
23	-0.18	-0.19	0.19	26.80
22	-0.28	-0.29	0.09	25.78
21	-0.37	-0.40	-0.02	24.60
20	-0.47	-0.51	-0.13	23.50
19	-0.58	-0.62	-0.24	22.40
18	-0.68	-0.73	-0.35	21.22
17	-0.79	-0.85	-0.47	20.00
16	-0.90	-0.97	-0.59	18.90
15	-1.01	-1.09	-0.71	17.78
14	-1.13	-1.22	-0.84	16.55
13	-1.25	-1.35	-0.97	15.34
12	-1.37	-1.49	-1.11	14.17
11	-1.51	-1.63	-1.25	13.00
10	-1.65	-1.79	-1.41	11.71
09	-1.80	-1.95	-1.57	10.43
08	-1.96	-2.13	-1.75	9.33
07	-2.14	-2.33	-1.95	8.06
06	-2.34	-2.54	-2.16	6.90
05	-2.57	-2.79	-2.41	5.70
04	-2.84	-3.07	-2.69	4.56
03	-3.17	-3.42	-3.04	3.39
02	-3.62	-3.89	-3.51	2.24
01	-4.35	-4.64	-4.26	1.12
00	---	---	---	---

As can be seen from Table 4, Total Raw Scores of 49 and 50 on Level I would be above ceiling on Level H and were, therefore, inestimable. Similarly, a Total Raw Score of 0 on Level I could not be scaled to Level H. The standard error of the Rasch equating was estimated from the translation constant. G_{HI} had a standard error of about $3.5/(NK)^{1/2}$ logits, where N is the number of persons taking each form and k is the number of items in the revised anchor test. The standard error for $G_{HI} = 0.38$, was, therefore, about 0.13 logits. In raw score units, this indicated that the error of the Rasch equating was approximately two points.

Tabulating the Equipercentile equating. As previously mentioned, ETS supplied vertically equated Total Raw Scores for Level I in terms of Level H. These estimates are reported in the "Equated I←H Equipercentile" column of Table 5. As with the Rasch method, Total Raw Scores of 49 and 50 on Level I were above ceiling on Level H; Total Raw Scores of 0-6 were given the same equated values. Therefore, it was only possible to evaluate the results based upon Total Raw Scores between 6 and 48.

The standard error of the Equipercentile equating was approximately equal to the standard error of measurement (SE_M) for Levels H and I. The SE_M corresponding to these tests was approximately 2.5 raw score points.

Rasch v. Equipercentile equating. In Table 5, the equated I←H Total Raw Scores based on Equipercentile v. Rasch methods are shown for each Total Raw Score on Level I. These values are plotted in Figure 2.

From Figure 2, it can be seen that the equated Total Raw Scores on Level I in terms of Level H were consistently higher based upon the Equipercentile method at or above a Total Raw Score of 13 on Level I. This means that, at or above a chance level score, a student would be at a relative advantage to take the higher, out-of-level Form I. However, while systematic,

TABLE 5.-- Equated I⇽H Total Raw Scores based on Equipercentile and Rasch methods for each Total Raw Score on Level I

TOTAL RAW SCORE$_I$	EQUATED I⇽H EQUIPERCENTILE	EQUATED I⇽H RASCH
50	---	---
49	---	---
48	50.00	48.88
47	49.50	48.24
46	49.00	47.61
45	48.33	47.00
44	47.83	46.25
43	47.33	45.50
42	46.80	44.77
41	46.20	44.00
40	45.50	43.16
39	45.00	42.39
38	44.33	41.56
37	44.00	40.71
36	43.00	39.79
35	42.00	38.92
34	41.50	38.00
33	40.50	37.00
32	40.00	36.08
31	39.00	35.09
30	38.00	34.09
29	37.00	33.18
28	36.00	32.10
27	35.00	31.00
26	34.00	30.00
25	33.00	29.00
24	31.00	27.90
23	30.00	26.80
22	28.00	25.78
21	27.00	24.60
20	26.50	23.50
19	25.00	22.40
18	24.00	21.22
17	22.00	20.00
16	20.00	18.90
15	18.00	17.78
14	17.00	16.55
13	15.00	15.34
12	14.33	14.17
11	13.00	13.00
10	12.00	11.71
09	11.33	10.43
08	10.00	9.33
07	9.00	8.06
06	8.00	6.90
05	---	---
04	---	---
03	---	---
02	---	---
01	---	---
00	---	---

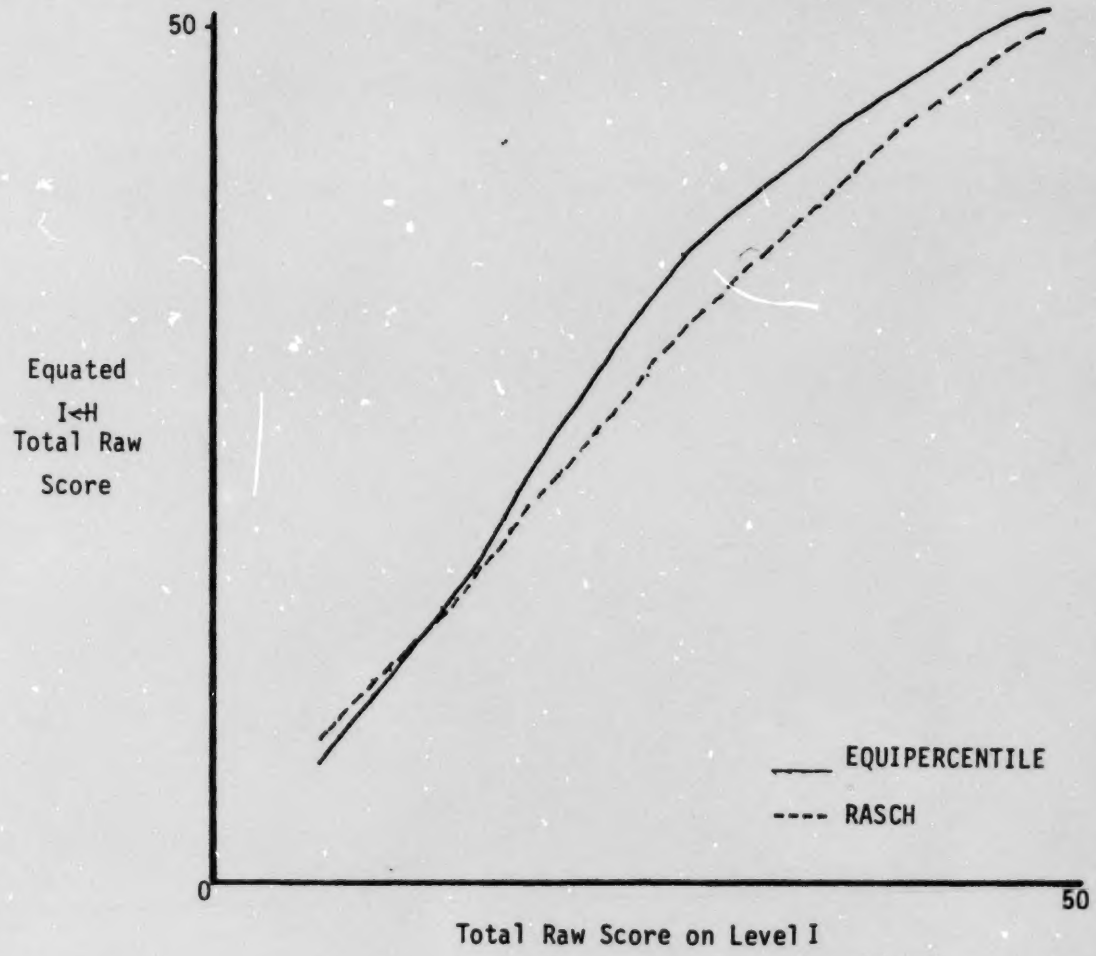

Fig.2.-- Plot of equated I←H Total Raw Scores based on Equipercentile v. Rasch methods for each Total Raw Score on test Level I .

the differences between the two methods were <u>not</u> statistically significant (χ^2_{13}=16.45). At any given Total Raw Score on I, the equated H←I scores using either method were within two standard errors of each other. Therefore, it may be concluded that the Rasch and Equipercentile methods estimated invariant vertically equated scores. Possible reasons for their discrepancies will be discussed in the "Conclusions and Implications" section of this paper.

CONCLUSIONS AND IMPLICATIONS

The purpose of this study was to investigate whether a test publisher's vertically scaled scores based on Equipercentile equating would be similar to the results of a school district whose scores would be equated by the Rasch method. For this study, two random samples of 100 sixth grade students from a suburban New York school district were drawn, one taking Level H of the STEP reading comprehension test, the other taking out-of-level Form I. The original anchor test vertically linking these tests was then revised by checking assumptions for Rasch equating. The revised anchor test, which met the assumptions of the Rasch equating method, was then used to estimate equated Total Raw Scores on Level I in terms of Level H. These equated scores based on the Rasch method were then compared to those supplied by the test publisher, who used the Equipercentile method. The results showed that, while the two sets of equated scores were statistically invariant, the Equipercentile method yielded systematically higher equated scores on Level I in terms of Level H for every Total Raw Score on Level I that was at or above chance level.

The implication of this study was that, based on the Rasch equating, a given student in the local school district might marginally benefit from taking the out-of-level Form I. However, this result was limited by two important issues related to the vertical equating methods.

First, it was clear from this study that, when using the Rasch procedure, assumptions must be carefully evaluated from the original anchor test to assure that the data fit the stringent criteria of this method. However, instead of proceding to average the differences between the difficulties of all anchor items in the original anchor test, it would seem advisable to initially delete items which were either evidently miskeyed or which became relatively harder at the higher test level. This initial editing procedure would eliminate the tendency of such items to underestimate the original translation constant which would serve as a basis for revising the anchor test. It was evident that this "two-phased" editing procedure would have, in this study, resulted in greater correspondence between the equatings estimated from the two methods.

A second limitation of the findings dealt with the assumptions of the Equipercentile method. Because the characteristics of the local school district were clearly not representative of the national sample upon which the publisher equated the test, there is no way to determine the effect of this problem. Sample dependent techniques require more careful matching than was possible for this study.

REFERENCES

Cook, L.L. & Douglass, J.B. Analysis of fit and vertical equating with the three-parameter model. Paper presented at the annual meeting of the American Educational Research Association, New York City, 1982.

Guskey, T.R. Comparison of a Rasch model scale and the grade equivalent scale for vertical equating of test scores. Occasional Paper, Office of Educational Research, University of Kentucky, 1979.

Holland, P.W. & Rubin, D.B. Test equating. New York: Academic Press, 1982.

Rentz, R. & Lenke, J. A comparison of Rasch and Equipercentile methods for vertical equating. Paper presented at the annual meeting of the American Educational Research Association, New York City, 1982.

Slinde, J.A. & Linn, R.L. Vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 15, 23-35.

Sontag, L.M. Vertical equating methods: A comparative study of their efficacy. Unpublished doctoral dissertation, Teachers College, Columbia University, 1984.

Wright, B.D. & Mead, R.J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, The University of Chicago, 1976.

Wright, B.D. & Stone, M.H. Best test design. Chicago: MESA Press, 1979.